Customer Churn Prediction-CRISP DM Data Report

Project title: Customer Churn Prediction

Date: 17 Dec 2025

1: Business Understanding

Problem Statement

SyriaTel is experiencing customer churn, which results in revenue loss and increased costs associated with acquiring new customers. The company needs a predictive approach to identify customers who are likely to churn so that retention actions can be taken proactively.

Business Objective

The objective of this project is to build a classification model that predicts whether a customer will churn based on historical customer behavior and service usage data.

Stakeholder

The primary stakeholders are SyriaTel’s customer retention and marketing teams, who can use churn predictions to design targeted retention strategies.

Success Criteria

The project is considered successful if:

- The model performs better than a baseline classifier
- Recall for churned customers is reasonably high
- Results are interpretable and actionable for business use

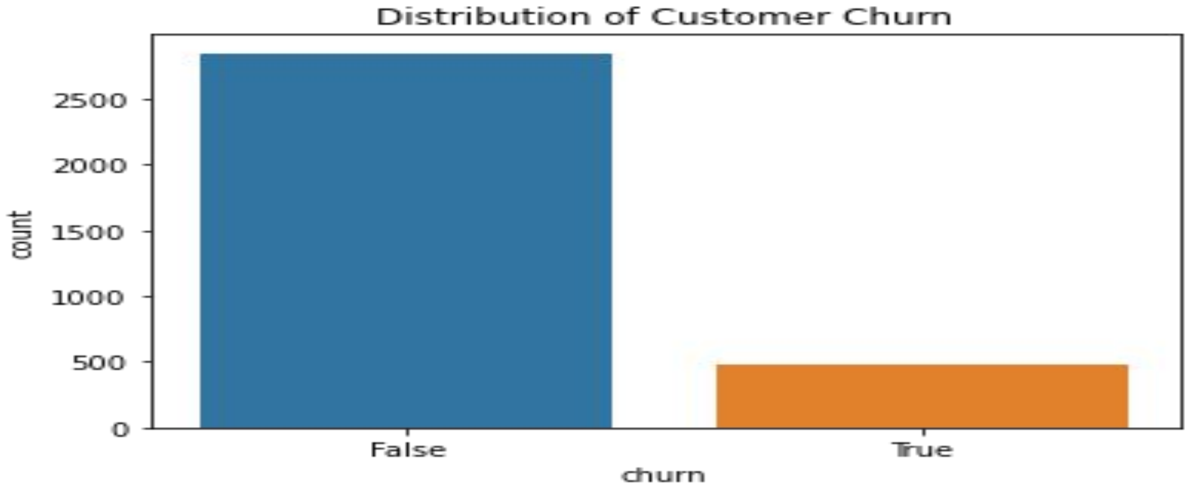
2: Data Understanding

The dataset contains customer-level information such as usage patterns, service plans, and account details. The target variable, `churn`, indicates whether a customer discontinued the service.

Churn Distribution

A churn distribution plot is used to examine class balance.

Figure 1: Distribution of Customer Churn



Interpretation:

The plot shows that non-churned customers significantly outnumber churned customers, indicating class imbalance. This suggests that accuracy alone is not an appropriate evaluation metric, and recall for the churn class should be emphasized during model evaluation.

3: Data Preparation

The data is prepared by separating predictors from the target variable and splitting the dataset into training and testing sets using stratified sampling to preserve the churn distribution. Numerical features are scaled, and categorical features are encoded using one-hot encoding.

All preprocessing steps are implemented within pipelines to prevent data leakage and ensure consistency across models.

4: Modelling

A baseline Logistic Regression model is developed to establish initial performance. This model serves as a benchmark for comparison with more complex models. Feature preprocessing and modeling are combined into a single pipeline for reproducibility and robustness.

A Decision Tree classifier is later introduced as a nonparametric alternative to capture nonlinear relationships in the data.

5: Evaluation

Model performance is evaluated using multiple complementary metrics.

Classification Report

The classification report summarizes precision, recall, and F1-score for each class.

Figure 2: Classification Report – Logistic Regression

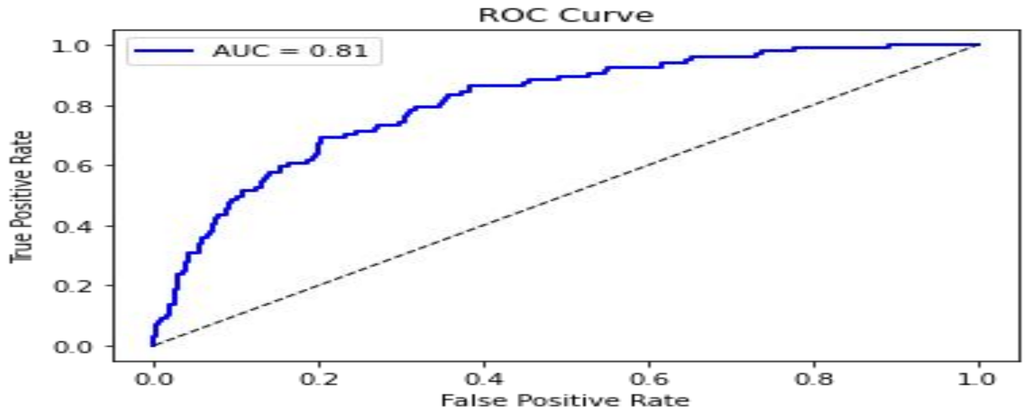
	precision	recall	f1-score	support
False	0.88	0.96	0.92	570
True	0.53	0.25	0.34	97
accuracy			0.86	667
macro avg	0.71	0.61	0.63	667
weighted avg	0.83	0.86	0.84	667

Interpretation:

The model performs very well in identifying non-churn customers, with high precision and recall. However, recall for churned customers is lower, meaning some customers who churn are not being correctly identified. This limitation motivates further model tuning and comparison with alternative models.

Figure 3: ROC Curve and AUC

The ROC curve evaluates model performance across different classification thresholds.



Interpretation:

The ROC curve shows that the model performs better than random guessing. The AUC score indicates moderate discriminatory power, suggesting the model can distinguish between churned and non-churned customers, though there is room for improvement.

6: Final Model Selection

Models are compared based on recall for the churn class, overall balance of metrics, and interpretability. The final model is selected based on its ability to identify churned customers while maintaining reasonable overall performance.

7: Findings, Limitations, and Recommendations

Key Findings

- Customer churn is a minority class, requiring careful metric selection
- Logistic Regression provides a strong, interpretable baseline
- Model performance is limited by class imbalance

Limitations

- The dataset does not include external factors such as competitor activity
- Class imbalance impacts the model’s ability to detect churn

Recommendations

- Use the model to flag high-risk customers for retention campaigns
- Explore class-weighted models or resampling techniques
- Incorporate additional customer behavior data in future models

8: Conclusion

This project demonstrates how classification modeling can support customer retention strategies by identifying customers at risk of churning. The insights generated can help SyriaTel make proactive, data-driven decisions to reduce customer attrition.

9: Next Steps

- Improve churn detection accuracy
- Apply advanced models
- Add more behavioral data
- Regularly update the model